ORIGINAL ARTICLE

Towards a predictive model for prevention of the risk of COVID-19 infection

*Djamel Toudert**
Department of Urban and Environmental Studies, El Colegio de la Frontera Norte, Baja California, Mexico

Abstract

Introduction: *The scarcity of person-centered applications aimed at developing awareness on the risk posed by the COVID-19 pandemic, stimulates the exploration and creation of preventive tools that are accessible to the population.* **Objective:** *To develop a predictive model that allows evaluating the risk of mortality in the event of SARS-CoV-2 virus infection.* **Methods:** *Exploration of public data from 16,000 COVID-19-positive patients to generate an efficient discriminant model, evaluated with a score function and expressed by a self-rated preventive interest questionnaire.* **Results:** *A useful linear function was obtained with a discriminant capacity of 0.845; internal validation with bootstrap and external validation, with 25 % of tested patients showing marginal differences.* **Conclusion:** *The predictive model with statistical support, based on 15 accessible questions, can become a structured prevention tool.*

KEY WORDS: *Predictive model of mortality. COVID-19. Preventive tool. Mexico.*

Hacia un modelo predictivo de carácter preventivo del riesgo de infección por COVID-19

Resumen

Introducción: *La escasez de aplicaciones centradas en la persona y con vistas al desarrollo de la conciencia del riesgo que representa la pandemia de COVID-19 estimula la exploración y creación de herramientas de carácter preventivo accesibles a la población.* **Objetivo:** *Elaboración de un modelo predictivo que permita evaluar el riesgo de letalidad ante infección por el virus SARS-CoV-2.* **Métodos:** *Exploración de datos públicos de 16 000 pacientes positivos a COVID-19, para generar un modelo discriminante eficiente, valorado con una función score y que se expresa mediante un cuestionario autocalificado de interés preventivo.* **Resultados:** *Se obtuvo una función lineal útil con capacidad discriminante de 0.845; la validación interna con bootstrap y la externa, con 25 % de los pacientes de prueba, mostraron diferencias marginales.* **Conclusión:** *El modelo predictivo, basado en 15 preguntas accesibles puede convertirse en una herramienta de prevención estructurada.*

PALABRAS CLAVE: *Modelo predictivo de la letalidad. COVID-19. Herramienta preventiva. México.*

Correspondence:
*Djamel Toudert
E-mail: toudert@colef.mx

Date of reception: 02-09-2020
Date of acceptance: 02-02-2021
DOI: 10.24875/GMM.M21000551

Gac Med Mex. 2021;157:231-236
Contents available at PubMed
www.gacetamedicademexico.com

Introduction

Exacerbation of the crisis caused by the COVID-19 pandemic generated an important demand for models predictive of the risk after SARS-CoV-2 infection.[1] Within the framework of this effort, predictive models were developed intended to support planning and resource management processes, refine medical protocols, and outline collective prevention policies.[1,2] However, predictive modeling has hardly focused on people aiming at generating higher involvement in conscious individual prevention. The latter plays a significant role in collective, informed and, especially, collaborative prevention structuring.[3]

The purpose of this research was to explore the universe of public data available on COVID-19 in Mexico,[4,5] in order to generate a predictive model of case fatality rate and its implementation in a questionnaire with accessible questions to the general population. Modeling was developed with a discriminant function of COVID-19-positive cases, through the outcome of recovering or dying, with each one of the predictive variables being given a score by means of the scoring technique.[6,7] To provide this tool with greater dissemination, the resulting questionnaire can be converted into an online-available application (app) where, in addition, personalized preventive advice is displayed based on the score obtained.

Methods

This study used publicly-open data on COVID-19 accumulated up to July 26, 2020, published and updated on a daily basis by the General Directorate of Epidemiology of the Ministry of Health of the federal government.[4] Open data from the National Institute of Statistics and Geography COVID-19 analytical viewer were also considered.[5]

Eight thousand patients who remained alive for more than two months after having tested positive for SARS-CoV-2 and 8,000 who were positive for the same infection and who died were randomly selected from the Ministry of Health database.[4] The municipalities of residence and variables corresponding to the 16,000 analyzed patients were retrieved from the National Institute of Statistics and Geography database.[5]

Taking into account that most of the analyzed data are expressed in categories, homogenization was carried out through a segmentation into four incremental classes by means of the scalable means algorithm.[8] The resulting data from 75 % of the positive patients were used to calculate the basic discriminant model, and those of the remaining 25 %, to carry out the model external validation process.

Of the set of variables explored owing to their thematic and epistemological proximity to the study, 15 variables were selected for their significant discriminant nature (chi-square test with $p < 0.05$) with regard to the outcomes of recovering or dying (Table 1).[9,10]

Subsequently, a multiple correspondence analysis was carried out, with all 27 axes (73.67 % of total explained variance) that were significant being preserved ($p < 0.05$).[6,11] With these factors, Fisher's discriminant linear function was calculated, which allowed generating the prediction model, followed by attribution of coefficients to the categorical variables using a score function (Table 2).[7,12] The score distribution defined the prediction of three situations:

- The optimal prognosis of overcoming the disease.
- A prediction of alert for patients likely to die.
- Indecision, which groups positive individuals with an intermediate score.

Evaluation of the developed model was carried out by means of internal validation with the bootstrap technique with 1,000 resamplings,[13,14] which was deepened with the support of the receiver operating characteristic (ROC) curve and the LIFT chart.[15] External validation was also carried out with 25 % of positive individuals, a population that did not participate in the generation of the model.

Results

Evaluation of the prediction model quality and efficiency was carried out in two complementary steps: internal validation and external validation.

Internal validation of the model

Validation of the model internal stability shows minimal variations between the result of the model basic calculation and the results with bootstrap.[16] In all cases, those differences were lower than 0.2 %, which indicated good internal stability. In the same vein, the ROC curve outlined a cutoff score of 575, which characterizes a sensitivity of 76.6 % and specificity of 23.1 %, both acceptable within the framework of the objectives established for this investigation (Fig. 1).

Table 1. Demographic, clinical and contextual characteristics of the analyzed population*

Characteristics	Alive (%)	Deceased (%)	Total (n = 16,000)
V1. Gender			
Males	26.11	32.41	9363
Females	23.89	17.59	6637
V2. Age groups			
1-4	0.30	0.08	60
5-14	0.57	0.04	98
15-24	3.55	0.24	606
25-34	11.61	1.37	2076
35-44	12.38	4.18	2649
45-64	17.13	22.10	6276
≥ 65	4.47	22.00	4235
V3. Pregnancy	90.00	10.00	70 (1.05)
V4. Obesity	43.87	56.13	3476 (21.72)
Comorbidity**			
V5. Diabetes	24.90	75.10	4061 (25.38)
V6. Chronic obstructive pulmonary disease	22.25	77.75	445 (2.78)
V7. Immunosuppression	29.88	70.12	328 (2.05)
V8. Hypertension	26.37	73.63	4733 (29.58)
V9. Cardiovascular	23.93	76.07	560 (3.5)
V10. Chronic kidney failure	15.73	84.27	674 (4.21)
V11. Other	34.62	65.38	621 (3.88)
V12. Positive patients by municipality			
1-948	14.40	16.08	4876
949-3,228	15.19	13.61	4608
3,229-6,832	14.01	14.39	4543
6,833-1,1587	6.40	5.91	1970
V13. 3 to 5-year population attending school by municipality			
0-49.36	2.68	3.44	979
49.37-61.76	15.74	16.32	5129
61.77-74.69	23.20	23.16	7417
74.70-100	8.36	7.06	2466
V14. Medical sector of care provision (population: 15,894)			
State	1.13	1.15	362
IMSS	15.11	27.46	6765
ISSSTE	1.86	3.72	888
Pemex	0.69	0.75	229
Municipal	0.04	0.04	13
University	0.03	0.03	9
Private	1.77	0.58	374
Sedena	0.38	0.49	138
Semar	0.52	0.23	118
Ministry of Health	28.52	15.50	6997
V15. Born in a State different from place of residence	10.99	13.06	3848 (24.05)

*All variables in the table are significant; $p < 0.05$.
**Only the affirmative modality of the comorbidity, pregnancy and obesity variables was included in the table, given that negation equals zero. IMSS (*Instituto Mexicano del Seguro Social*) = Mexican Institute of Social Security; ISSSTE (*Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado*) = Institute of Social Security and Services for State Workers; Pemex = Petróleos Mexicanos; Sedena (*Secretaría de la Defensa Nacional*) = Ministry of National Defense; Semar (*Secretaría de Marina*) = Ministry of the Navy.

The discriminant capacity of the model by means of the area under the curve (AUC) was 0.845, with a 95 % confidence interval, which indicates that the prediction distinguishes between recovered and deceased subjects with a likelihood of 84.5 %. A model with this AUC value is considered useful for certain uses according to Swets's criteria,[17] while it is characterized as excellent according to Hosmer and Lemeshow criteria.[18]

External validation of the model

The comparison of the results obtained with the basic model and the calculation made with 25 % of the study patients suggested marginal differences (Table 3), which were contained within an absolute value range lower than 0.65 %, which does not appear to significantly affect the model efficiency. This can be observed in the LIFT chart, which reveals a model able to determine 76.57 % of correct predictions with 50 % of positive patients (Fig. 1).

The result of the discriminant function with categorical variables was used to develop a score function; each one of the positive patients was assigned to three possible zones based on the score they obtained. In our case, the high scores corresponded to the alert zone (622-1,000) and the low scores coincided with the optimal zone (0-528). The indecision zone identified individuals with scores ranging from 528 to 622, who could not be classified in any of the previous two (Table 4).

Discussion

Through discriminant modeling, the possibility of predicting recovery and death when contracting the SARS-CoV-2 virus in Mexico was established. The prediction model that was generated has the advantage of being supported by accessible public data; it consists of predictive variables that are easy to answer and, therefore, susceptible to facilitating the creation of a computational tool for preventive purposes. During the COVID-19 crisis, Wynants et al.[1] documented nearly 145 predictive models, 34 % of them addressing the mortality risk prognosis or prediction of the need for intensive care. Most models that showed advances in terms of knowledge about the disease[2,19] were considered to be biased because of a non-representative selection of control patients and model overfitting.[1]

Tabla 2. **Fisher's discriminant linear function internal and external validation**

Model development	Basic calculation		Calculation with bootstrap*	
Discriminated groups	**Well classified**	**Wrongly classified**	**Well classified**	**Wrongly classified**
Alive positives	4,492	1,508	4,481.50	1,518.50
n %	74.85	25.13	74.69 [0.63]	25.31 [0.63]
Deceased positives	4,761	1239	4,751.61	1,248.39
n %	79.35	20.65	79.19 [0.67]	20.81 [0.67]
Total	9,253	2,747	9,233.11	2,766.89
n %	77.11	22.89	76.94 [0.37]	23.06 [0.37]
Model verification	**Basic calculation**			
Alive positives	1,502	498		
n %	75.10	24.90		
Deceased positives	1,574	426		
n %	78.70	21.30		
Total	3,076	924		
n %	76.90	23.10		

*1,000 random resampling.
Standard deviation indicated in square brackets.

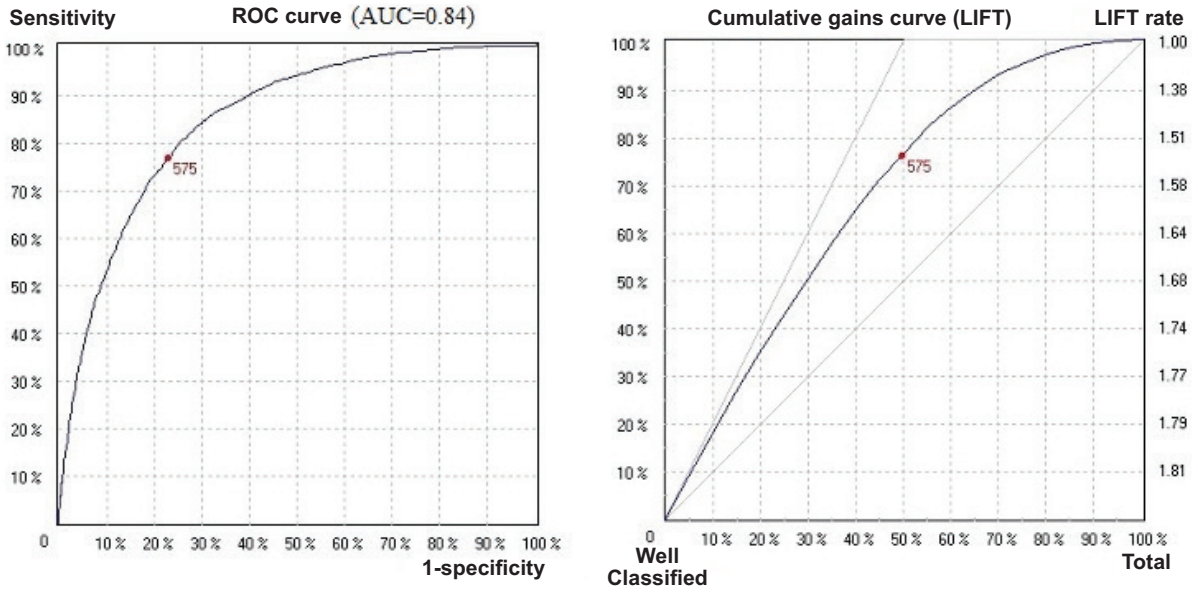

Figure 1. *Quality and efficiency of the predictive model.*

Although the present study does not reproduce the aforementioned methodological deficiencies, the proposed model does not appear to be an appropriate instrument for making decisions of operational significance in the medical field. Even with an AUC of 0.845, the proposed model would acquire more efficiency with the incorporation of symptomatic and laboratory variables.[2] However, this would also distance the model from its preventive objectives induced by the ability of a common person to instantly answer a

Table 3. Discriminant and transformed coefficients of involved modalities

Characteristics	Discriminant function coefficients	Transformed coefficients
V1. Gender		
Males	1.936	15.85
Females	−2.731	0
V2. Age groups (years)		
1-4	-21.688	29.5
5-14	−28.67	5.78
15-24	−30.371	0
25-34	−29.564	2.74
35-44	−18.279	41.08
45-64	4.641	118.95
≥ 65	24.365	185.96
V3. Pregnancy	3.323	20.47
V4. Obesity	1.918	46.73
Comorbidity		
V5. Diabetes	Municipal	-10.005
V6. Chronic obstructive pulmonary disease	−1.38	8.66
V7. Immunosuppression	3.735	24.7
V8. Hypertension	3.981	19.37
V9. Cardiovascular	1.918	46.73
V10. Chronic kidney failure	13.754	72.78
V11. Other	2.487	9.09
V12. Positive patients by municipality		
1-948	3.67	18.6
949-3,228	−1.735	0.24
3,229-6,832	14.01	14.39
6,833-11,587	−0.901	3.07
V13. 3 to 5-year population attending school by municipality		
0-49.36	6.038	27.46
49.37-61.76	0.297	7.96
61.77-74.69	−0.327	5.84
74.70-100	8.36	7.06
V14. Medical sector of care provision		
State	1.718	325.81
IMSS	8.854	350.06
ISSSTE	6.917	343.47
Municipal	−10.005	285.99
University	−18.09	258.52
Pemex	−8.596	290.77
Private	−21.188	247.99
Sedena	5.287	337.94
Semar	−9.127	288.97
Ministry of Health	−8.069	292.56
V15. Born in a State different from place of residence		
Yes	1.065	4.76
No	−0.337	0

IMSS (*Instituto Mexicano del Seguro Social*) = Mexican Institute of Social Security; ISSSTE (*Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado*) = Institute of Social Security and Services for State Workers; Pemex = Petróleos Mexicanos; Sedena (*Secretaría de la Defensa Nacional*) = Ministry of National Defense; Semar (*Secretaría de Marina*) = Ministry of the Navy.

Table 4. Distribution of patients with a 10 % tolerated error rate

Positive patients	Optimal zone (0-528 points)		Indecision zone (528-622 points)		Alert zone (622-1,000 points)	
	n	%	n	%	n	%
Recovered	4,777	59.71	2,423	30.29	800	10
Deceased	807	10.09	2,992	37.40	4,201	52.51
Total	5,584	34.90	5,415	33.84	5,001	31.26

set of questions. Taking into account the scarcity of individual prognostic tools,[1] the dilemma in this study was solved in favor of a preventive scope.

In theoretical terms, the model appears to corroborate, as identified in other investigations, the importance of age, gender and comorbidity for mortality risk.[1,3] As a specific contribution of this study, the weight of contextual factors is underlined, including the sectors that care for patients with COVID-19, the number of positive patients and the rate of three- to five-year-old children who attend school by municipality of residence. The explanation for the incidence of the first two variables seems obvious, but in the case of the latter, the closest is a hypothetical referent of a sort of social immunity in that category of children.[20]

In practical terms, the developed model exhibits statistical qualities that enable for it to become a preventive app containing 15 questions structured based on the variables shown in table 3. Within the framework of this app, a score (table 3 transformed coefficient) will be assigned to each one of the answer modalities, which will be added once the questionnaire is completed to then assign the case to one of the three decision zones (alert, indecision and optimal). Assignment of consultation to a decision zone could reach a level of usefulness accompanied by personalized preventive medical advice.[21]

Conclusion

A predictive model was obtained with a degree of usefulness that has statistical and nomological support, which allows its conversion into a preventive app. A preliminary example of this app can be found at https://bit.ly/2YPlox8.

Conflict of interests

The authors declare that they have no conflicts of interest.

Funding

The authors did not receive any sponsoring to carry out this article.

Ethical disclosures

Protection of human and animal subjects. The authors declare that no experiments were performed on humans or animals for this research.

Confidentiality of data. The authors declare that no patient data appear in this article.

Right to privacy and informed consent. The authors declare that no patient data appear in this article.

References

1. Wynants L, van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ. 2020;369:m1328.
2. Ciceria F, Castagnaa A, Rovere-Querinia P, de Cobellia F, Ruggerib A, Gallib L, et al. Early predictors of clinical outcomes of COVID-19 outbreak in Milan, Italy. Clin Immunol. 2020;217:108509.
3. Liu T, Liang W, Zhong H, He J, Chen Z, Guanhao H, et al. Risk factors associated with COVID-19 infection: a retrospective cohort study based on contacts tracing. Emerg Microbes Infec. 2020;9:1546-1553.
4. Dirección General de Epidemiología, Secretaría de Salud; [Internet]. Mexico: Datos abiertos; 2020.
5. Instituto Nacional de Estadística y Geografía [Internet]. Mexico: Visualizador analítico para el COVID-19; 2020.
6. Lebart L, Morineau A, Piron M. Statistique exploratoire multidimensionnelle. France: Dunod; 2000.
7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-138.
8. Bahmani B, Moseley B, Vattani A, Kumar R, Vassilvitskii S. Scalable k-means++. Proceedings of the VLDB Endowment. 2012;5:622-633.
9. Morineau A. Note sur la caractérisation statistiques d'une classe et les valeurs-tests. Bull Techn Centre Statist Inform Appl. 1984;2:20-27.
10. Nakacha JP, Confais J. Approche pragmatique de la classification. France: Technip; 2004.
11. Bardos M. Analyse discriminante. Application au risque et scoring financier. France: Dunod; 2001.
12. Droesbeke JJ, Lejeune M, Saporta G, editores. Modèles statistiques explicative pour données qualitatives. France: Editions Technip; 2005.
13. Bleeker SE, Moll HA, Steyerberg EW, Donders, G, Derksen-Lubsen DE, Grobbee SE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003;56:826-832.
14. Wahl S, Boulesteix AL, Zierer A, Thorand B, van de Wiel MA. Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC Med Res Methodol. 2016;16:144.
15. Vuk M, Curk, T. ROC curve, lift chart and calibration plot. Metodoloski zvezki. 2006;3:89-108.
16. Xie J, Qiu Z. Bootstrap technique for ROC analysis: a stable evaluation of Fisher classifier performance. J Electron. 2007;24:523-527.
17. Swets JA. Measuring the accuracy of diagnosis system. Science. 1988;240:1285-1293.
18. Hosmer DW, Lemeshow S. Applied logistic regression. USA: John Wiley and Sons; 2000.
19. Bello-Chavolla O, Bahena-López J, Antonio-Villa N, Vargas-Vázquez A, González-Díaz A, Márquez-Salinas C, et al. Predicting mortality due to SARS-CoV-2: a mechanistic score relating obesity and diabetes to COVID-19 outcomes in Mexico. J Clin Endocrinol Metab. 2020;105:1-10.
20. Kloc M, Ghobrial RM, Kuchard E, Lewickie S, Kubiak J. Development of child immunity in the context of COVID-19 pandemic. Clin Immunol. 2020;217:108510.
21. Barajas-Ochoa A, Andrade-Romo J, Ramos-Santillán V. Retos para la educación médica en México en los tiempos del COVID-19. Gac Med Mex. 2021;157:254-257.